CUADERNOS.INFO Nº 56
Versión electrónica: ISSN 0719-367x
http://www.cuadernos.info
https://doi.org/10.7764/cdi.56.60093

Enviado: 29/03/2023 / Aceptado: 09/06/2023

Research trends in the control of hate speech on social media for the 2016–2022 time frame

Tendencias en la investigación para el control del discurso de odio en las redes sociales para el período 2016-2022

Tendências na pesquisa sobre o controle do discurso de ódio nas redes sociais para o período 2016-2022

Ana M. Sánchez-Sánchez, Universidad Pablo de Olavide, Sevilla, España (amsansan@upo.es)

David Ruiz-Muñoz, Junta de Andalucía, Sevilla, España (david.ruiz.m@juntadeandalucia.es)

Francisca J. Sánchez-Sánchez, Universidad Pablo de Olavide, Sevilla, España (fsansan@upo.es)

ABSTRACT | The growth in the number of social media users has resulted in a corresponding rise in the spread of hate speech on these platforms, leading to a growing, but little studied, problem. The bibliometric study aimed to examine the research trend and identify the most productive authors, the most active institutions, the leading countries and the most employed virtual hate speech control mechanisms by analyzing 576 relevant publications from the Scopus database published between 2016-2022. The findings showed an increase in publication and India as a leading country/region in research on virtual hate speech control mechanisms. Deep learning and natural language processing systems were identified as the most commonly used control mechanisms. Based on the results, it is recommended that future researchers focus on multidisciplinary collaboration and valid mechanisms for different languages. This paper provides a general overview of the current state of research in this field and serves as a guide for authors and institutions in their research and collaboration strategies.

KEYWORDS: Hate speech; social media; detection; machine learning; deep learning; natural language processing systems; bibliometric analysis.

FORMA DE CITAR

Sánchez-Sánchez, A.M., Ruiz-Muñoz, D. & Sánchez-Sánchez, F.J. (2023). Research trends in the control of hate speech on social media for the 2016–2022 time frame. *Cuadernos.info*, (56), 89-116. https://doi.org/10.7764/cdi.55.60093

RESUMEN | *El crecimiento del número de usuarios de las redes sociales ha conllevado el correspondiente aumento de la difusión del discurso de odio en estas plataformas, dando lugar a un problema creciente y poco estudiado. El estudio bibliométrico buscó examinar la tendencia de la investigación e identificar a los autores más productivos, a las instituciones más activas, a los países líderes y los mecanismos virtuales de control del discurso de odio más empleados mediante el análisis de 576 publicaciones relevantes de la base de datos Scopus publicadas entre 2016-2022. Los hallazgos mostraron un aumento de las publicaciones y que la India es el país líder en investigación sobre mecanismos virtuales de control del discurso de odio. El deep learning y el natural language processing systems fueron identificados como los mecanismos de control más empleados. El estudio sugiere que la investigación futura debería centrarse en la colaboración multidisciplinar y en mecanismos de control válidos para diferentes idiomas. El artículo proporciona una visión general del estado actual de la investigación en este campo y sirve de guía para autores e instituciones en sus estrategias de investigación y colaboración.*

PALABRAS CLAVES: *Discurso de odio; redes sociales; detección; aprendizaje automático; aprendizaje profundo; sistemas de procesamiento de lenguaje natural; análisis bibliométrico.*

RESUMO | O crescimento do número de usuários das redes sociais tem levado a um correspondente aumento da propagação do discurso de ódio nestas plataformas, dando origem a um problema crescente e pouco estudado. O estudo bibliométrico teve como objetivo examinar a tendência da pesquisa e identificar os autores mais produtivos, as instituições mais ativas, os países líderes e os mecanismos virtuais de controle do discurso de ódio mais utilizados, analisando 576 publicações relevantes da base de dados Scopus publicadas entre 2016-2022. Os resultados mostraram um aumento nas publicações e que a Índia é o principal país para pesquisas sobre mecanismos virtuais de controle do discurso de ódio. O *Deep Learning e Natural Language Processing Systems* foram identificados como os mecanismos de controle mais comumente usados. O estudo sugere que as pesquisas futuras devem se concentrar na colaboração multidisciplinar e nos mecanismos de controle válidos para diferentes idiomas. O documento fornece uma visão geral do estado atual da pesquisa neste campo e serve como um guia para autores e instituições em suas estratégias de pesquisa e colaboração.

PALAVRAS-CHAVE: Discurso de ódio; redes sociais; detecção; aprendizagem automática; aprendizado profundo; sistemas de processamento de linguagem natural; análise bibliométrica.

INTRODUCTION

The creation and dissemination of hate speech is becoming a substantial problem, which has led to the proposal of several international initiatives aimed to identify the problem and develop effective countermeasures. Sellars (2016) analyses different definitions in the academic and legislative sphere, identifying certain traits (the fact of addressing a group, or an individual as a member of a collective, the presence of content that expresses hatred, causes harm, incites wrongdoing, beyond the speech itself or the public nature of the speech) that, although they do not generate a single definition, increase confidence that the speech in question is worthy of being identified as hate speech. According to Fortuna and Nunes (2018), hate speech is "the content that promotes violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, veteran status, and sexual orientation/gender identity".

The growing problem generated by hate speech is manifested by the emerging literature on its detection (Burnap & Williams, 2016), with aggressive speech (Risch & Krestel, 2018), directed against certain ethnicities (Waseem, 2016), against women (Fersini et al., 2020), or against certain religions (Albadi et al., 2018), which is why it is so important to detect such hateful content.

With the Internet's widespread use, social networks have become essential tools in online communication and a space for hate speech and cyberbullying (Giumetti et al., 2022). The role of social media in relation to hate speech is complex and contradictory. On the other hand, social media platforms have policies in place to prohibit explicit hate speech, and they also provide a means for the spread of hateful messages. This creates a challenge for social media platforms, as they must balance freedom of expression with the need to address hate speech and prevent its spread (Ben-David & Matamoros Fernández, 2016).

Social media has generated a specific type of hate speech, cyberbullying, whereby an attempt is made to harm an individual or a collective through the use of digital media (Dadvar et al., 2015), and can occur via various modalities: sharing or posting multimedia offensive content without the owner's permission (Dewani et al., 2021), making automatic detection more difficult. However, cyberbullying via textual content is far more common. Automated hate speech detection is an important tool for detecting and preventing the spread of hate speech, especially on social media.

Social networks act as an amplifier of hate speech, so the quest for automatic mechanisms for hate speech recognition is becoming an important research topic that needs a comprehensive and multidisciplinary approach to analyze, detect, and successfully neutralize its negative impact (Ramírez-García et al., 2022).

Numerous approaches have been developed in recent years to automatically detect hate speech on social media, but much of it is still undetectable because it is not valid for different types of languages, communication (Omar & Hashem, 2022), and multimedia content. The availability of suitable quality data also remains a challenge for automatic detection and small datasets (Albadi et al.,2018).

Artificial intelligence methods and techniques, including machine learning (ML), deep learning (DL), transfer learning (TL), and recently pretrained language models (NLPS) have been an essential step to detect abusive content (Alrashidi et al., 2022).

TL is a type of ML that can be used to learn from data that has been previously learned by another ML algorithm. Pretrained methods have been playing a major role in driving the development of many ML and NLP areas including hate speech detection. Much of the existing work on abusive content detection, focuses on using supervised ML (Kanan et al., 2020). DL is based on artificial neural networks consisting of complex deep learning algorithms. Recent studies have found that the combination of two or more DL models outperforms the use of a single DL model (Al-Hassan & Al-Dossari, 2022). TL is a notion in the ML area in which prior knowledge learned is applied to solve a problem from a different subject and task that is connected in some way. Recently, TL approaches were applied in some studies for abusive content detection, such as Mozafari and colleagues (2020), highlighting specially the results obtained with BERT models.

Mishra (2021) conducted a descriptive study on the type of publications, research areas, affiliation, countries, and keywords related to hate speech from 1962 to 2021, although this research did not focus on social networks. Tontodimamma and colleagues (2021) expanded on his research by analyzing the basics of hate speech between 1992 and 2019 and highlighting the impact of social networks. Ramírez-García and collaborators (2022) took this research further by providing an updated analysis of previous studies and evaluating the interrelation between hate speech and social networks, finding that the topic gained importance in the scientific community starting in 2017. Our study adds to existing research by examining a specific aspect that has received little attention so far, the mechanisms for detecting hate speech on social networks, using interdisciplinary methods such as keyword co-occurrence analysis (Vargas-Quesada et al., 2017), paper production and citation analysis (Zamora-Bonilla & González de Prado, 2014), or co-citation analysis (Córdoba-Cely et al., 2012).

This study takes bibliometric analysis as a starting point, as it is considered to be a very effective tool, providing data and information that can be used researchers,

and influential groups interested in improving the quality of research or offering solutions to different problematic situations (Nandiyanto et al., 2020).

Tontodimamma and colleagues (2021) emphasize that automatically detecting and classifying hate speech using machine-learning strategies to correctly assess hate forms of online speech. The aim of our work is deepen research direction proposed by these authors, and try to answer the following research questions:

- What is the general distribution of publications by year, institutions, countries, and authors in the development and application of control mechanisms, and what collaborations have been established? What are the most cited publications in the field of control mechanisms?
- What are the most commonly used automatic mechanisms for detecting hate speech and how has their use evolved over time?
- What are the main keywords, consistencies, and research gaps in the field of control mechanisms?

MATERIAL AND METHODS

The two most commonly used databases for bibliometric studies are Scopus and Web of Science. Although both sources can provide the information needed for our analysis, Scopus was selected because of its greater coverage of journals and total number of citations, as well as its use in similar studies (Singh et al., 2021; Martín-Martín et al., 2021; Mishra, 2021; Ramírez-García et al., 2022).

Although this study is not a classic systematic review, it is a scientometric article that uses a rigorous analysis of the scientific literature. To ensure a clear and understandable methodology, the PRISMA guidelines (Page, 2021) were adapted for this study (figure 1).

We chose the 2016-to-2022-time frame for this study because 2016 was he first year in which relevant publications appeared on the topic of social media, recognition, and hate speech. Prior, there was no consensus on the definition of hate speech (Strossen, 2016; Sellars, 2016), so it was not possible to develop detection mechanisms. The year 2022 was used an upper bound to obtain complete annual data. The results are consistent with previous studies showing a growing interest in the basic descriptive metrics of scientific production on hate speech and social networks as of 2017 (Ramírez-García et al., 2022), and reflect the importance of addressing this issue by detecting, preventing, and punishing hate speech on social media which has been highlighted by various institutions and organizations (Fernández et al., 2015; Movement Against Intolerance, 2015; ECRI General Policy Recommendation N°15, 2015).

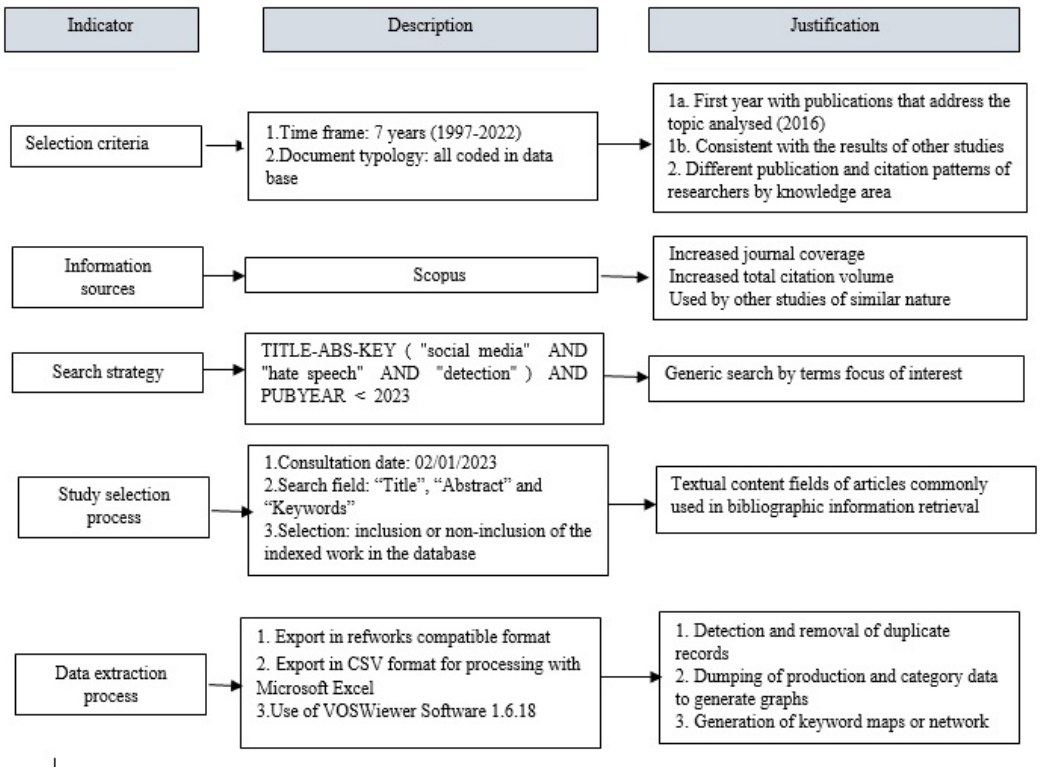

Figure 1. PRISMA diagram depicts data collection from Scopus database

Source: Own elaboration.

The search terms used were social media, detection, and hate speech. These could appear in the title, abstract, and keywords. Initially, 577 publications were found. To detect possible duplicates in the database, we used the bibliographic manager Refworks, which detected three possible duplicates. After verification, only one turned out to be true. Before analyzing the information, a standardization of the different elements (authors, institutions, publications) was performed.

RESULTS

Descriptive analysis

The growing number of social media users has led to a corresponding increase in the prevalence of hate speech on these platforms. This has led to a growing need for control mechanisms, which are the focus of policies and laws developed by international organizations such as the United Nations (2019), the European Commission (2020), and UNESCO (2021). Our study found that the number of publications addressing this issue increased significantly in the last years (2019-2022) of the period studied, coinciding with the implementation of these regulations and strategies.

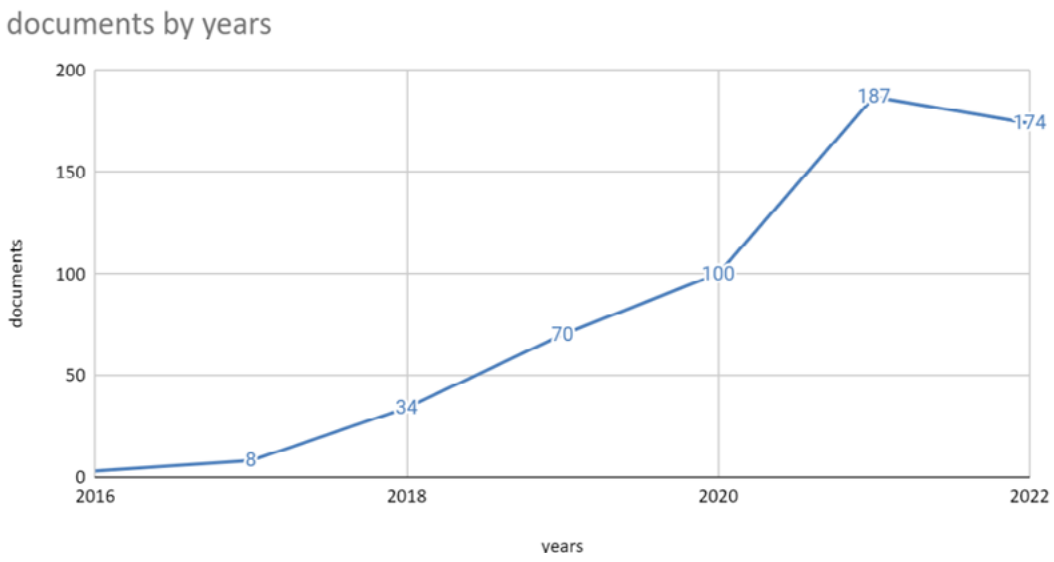

Figure 2. Annual evolution of publications

Source: Own elaboration.

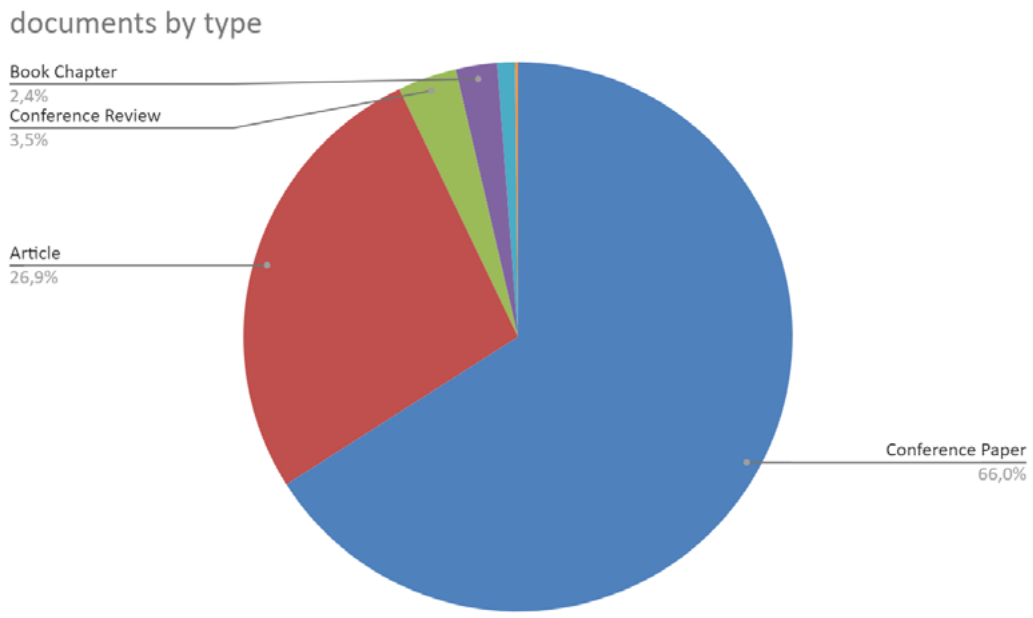

Figure 3. Documents Type

Source: Own elaboration.

The upward trend in the number of publications is broken in 2022, as 13 fewer documents are recorded than in 2021.

Of the 576 publications analyzed, most (573) were in English, while two were written in Spanish, and one, in Turkish.

The types of documents were distributed as in Figure 3.

	Technical Sciences		Social Sciences	
Control mechanisms	**Keywords (No.)**	**Keywords (%)**	**Keywords (No.)**	**Keywords (%)**
DL	144	14.89 %	40	15.44 %
ML	103	10.65 %	33	12.74 %
Classification (of information)	102	10.55 %	31	11.97 %
Natural language processing systems (NLPS)	97	10.03 %	38	14.67 %
Learning systems	82	8.48 %	23	8.88 %
Learning algorithms	57	5.89 %	16	6.18 %
Computational linguistics	55	5.69 %	25	9.65 %
Text processing	54	5.58 %	12	4.63 %
Support vector machines	50	5.17 %	11	4.25 %
Text classification	50	5.17 %		
Long short-term memory (LSTM)	48	4.96 %		
Sentiment analysis	46	4.76 %	15	5.79 %
Decision trees	43	4.45 %	15	5.79 %
Embeddings	36	3.72 %		

Table 1. Control mechanisms by category

Source: Own elaboration.

The most recurrent type of document in this research area is the conference paper, as it is a format commonly used by researchers in the field of computer science.

When focusing on thematic research areas , the most fruitful are: Computer Science (45.64%), Engineering (13.83%), Mathematics (9.34%), Social Sciences (8.37%), and Decision Sciences (6.74%) are the top four disciplines.

These five categories account for 83.88% of the papers examined in our study.

The first three categories are scientific and technical in nature, with 537 publications, while the next two are scientific and social in nature, with 153 documents. This significant difference in the number of documents is logical, since the first category contains articles on both the development and application of control mechanisms, while the second category deals only with application.

To elaborate table 1, Scopus was filtered by the five subject areas that make up the two categories shown. The keywords corresponding to the 15 most frequent detection mechanisms were selected. They were then filtered according to the two categories created (Technical Sciences and Social Sciences), with the corresponding frequency assigned to each control mechanism.

Control mechanism	b	
DL	13.20	GROUP 1
NLPS	11.50	
ML	7.70	GROUP 2
Classification (I&T)	6.30	
Learning algorithms	5.00	
Text Processing	3.60	GROUP 3
Support vector machines	3.30	
LSTM	3.20	
Computational Linguistics	3.00	
Sentiment Analysis	2.80	

Table 2. Trend in the use of detection mechanisms

Source: Own elaboration.

Control mechanisms used in both the social and technical categories include DL, ML, classification, and NLPS. In the technical category, only text classification, LSTM, and Embeddings are used. A time trend analysis (y=a+bx, where "y" represents the number of uses of the control mechanism, "x", the year of use, "a", the origin of the line, and "b", the slope) of the use of these mechanisms showed that as from 2018, the 10 most commonly used mechanisms can be divided into three groups with similar patterns of use over time, as determined by a linear regression model.

Figure 4 shows the trend in the use of the different control mechanisms. We have distinguished three groups according to the value of the slope (positive in all cases), with this trend increasing over the years. DL and NLPS are the mechanisms that show the greatest increase in their use over time.

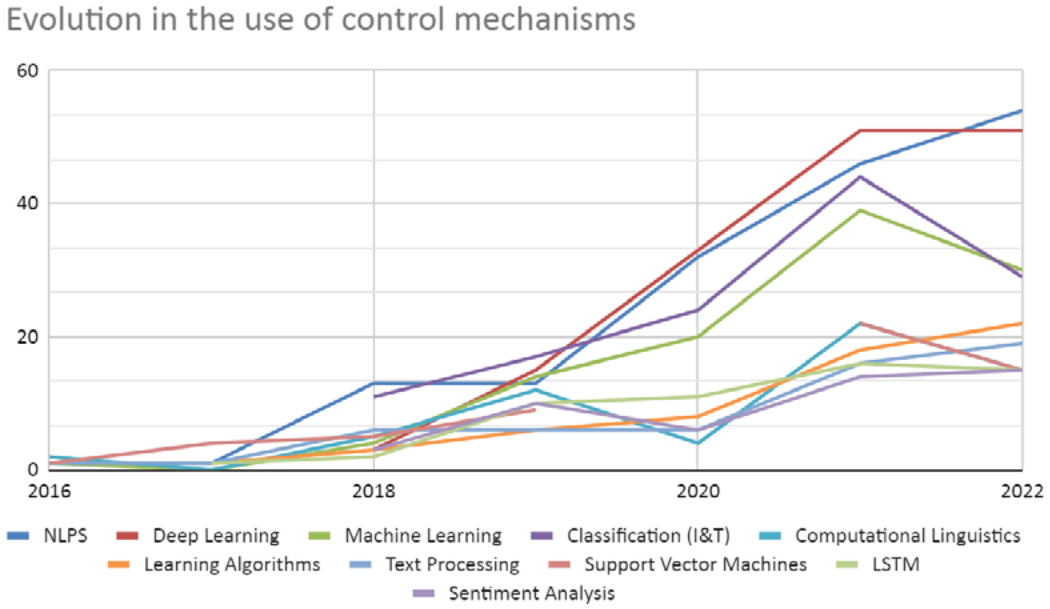

Figure 4. Evolution in the use of control mechanisms

Source: Own elaboration.

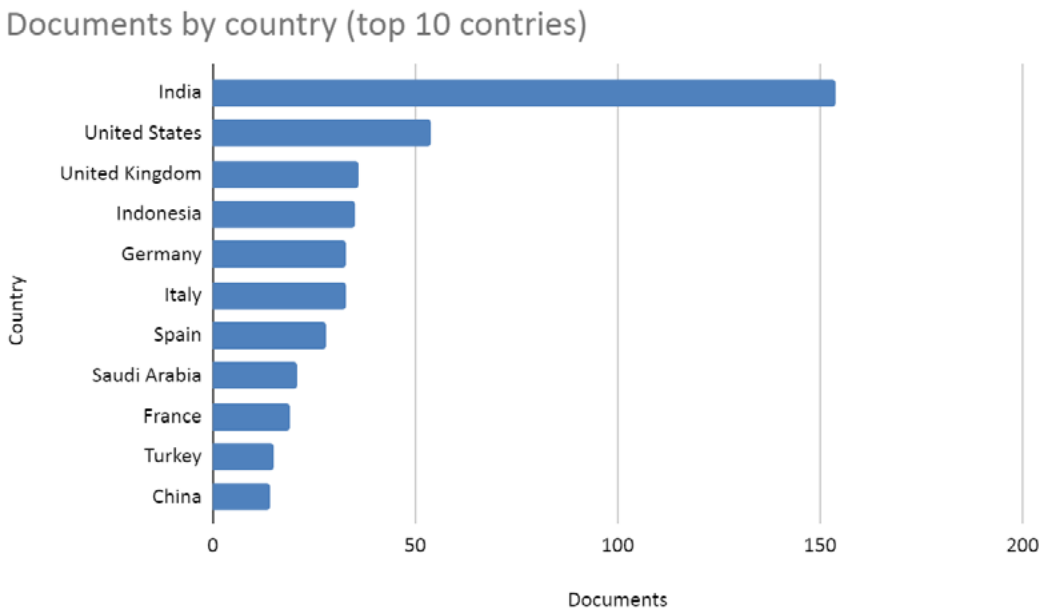

Figure 5. Top productive countries based on the number of publications

Source: Own elaboration.

India is the most productive country in terms of the number of publications, with 26.73% of the total scientific production studied.

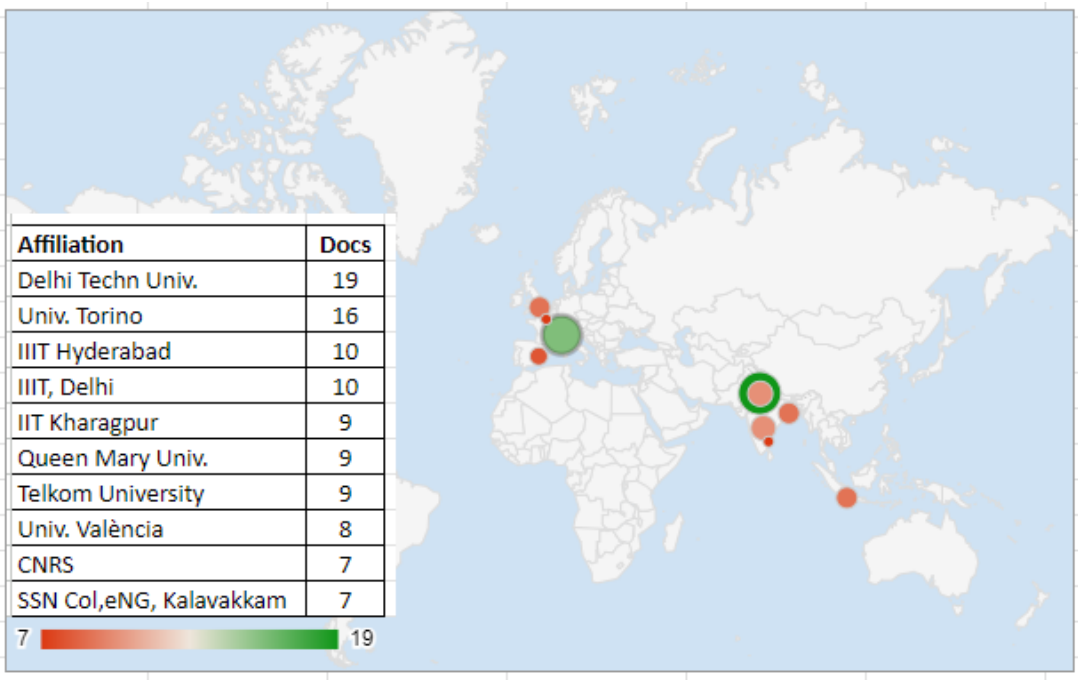

Figure 6. Top contributing institutions based on total publications

Source: Own elaboration.

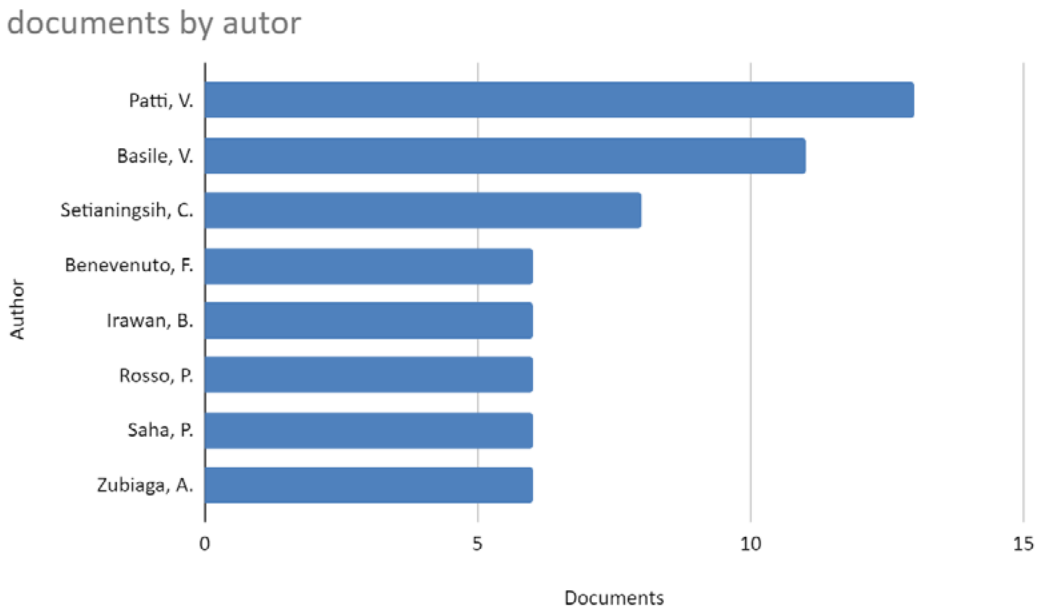

Figure 7. Top productive authors based on article count

Source: Own elaboration.

Out of the 10 most productive institutions, five are located in India.

The ranking of the most productive authors, with at least six publications, is shown in figure 7.

Only one of these authors (Saha), belongs to an Indian institution (I. I.T Kharagpur).

Regarding the most influential publications, the following can be considered as the most relevant if they have received at least 100 citations.

Analysis of the articles in table 3 shows that Twitter is the most commonly used social network for applying control mechanisms. Among these, NLP ranks first, and the linguistic model BERT, known for its advanced language processing capabilities, is the most commonly used within NLP.

Analysis of co-authorship

To determine the thresholds used in the various analyses performed with VOSviewer, the minimum values were set so as to lose as little information as possible about the relationships between the elements analyzed, while not to producing extensive lists and complex maps that are difficult to visualize and interpret.

In bibliometric analysis, co-authorship analysis is often used to examine various collaboration aspects. The resulting collaboration networks are created by analyzing co-authorship relationships (Glänzel & Schubert, 2004, p. 257; Romero & Portillo-Salido, 2019; Van Eck & Waltmann, 2020).

Document title	Authors	Cited by
Automated hate speech detection and the problem of offensive language	(Davidson et al., 2017)	873
Hateful symbols or hateful people? predictive features for hate speech detection on twitter	(Waseem & Hovy, 2016)	740
A Survey on Hate Speech Detection using Natural Language Processing	(Schmidt & Wiegand, 2017)	606
Predicting the type and target of offensive posts in social media	(Zampieri et al.,2019)	303
Hate speech detection: Challenges and solutions	(MacAvaney et al.,2019)	188
Us and them: identifying cyber hate on Twitter across multiple protected characteristics	(Burnap & Williams, 2016)	181
Hate me, hate me not: Hate speech detection on Facebook	(Del Vigna et al.,2017)	155
A measurement study of hate speech in social media	(Mondal et al.,2017)	124
Analyzing the targets of hate in online social media	(Silva et al.,2016)	124
A dataset of Hindi-English code-mixed social media text for hate speech detection	(Bohra et al., 2018)	107
Effective hate-speech detection in Twitter data using recurrent neural networks	(Pitsilis et al.,2018)	105
A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media	(Mozafari et al.,2020)	104

Table 3. Most-cited publications in subject of control mechanisms of virtual hate speech from 2016 to 2022

Source: Own elaboration.

By conducting a co-authorship analysis and taking authors as the unit of analysis, and only considering authors with at least 3 publications and 1 citation, a map is generated that allows us to analyse the temporal evolution of these collaborations.

The map shows five clusters of collaboration between the authors. The first one consists of Florio, Polignano, and Basile. They analyze the use of the BERT technique to monitor hate speech on Italian Twitter. The second is made up of Patti, Basile, and Pamungkas, who specialize in cross-linguistic classification and classification (of information) and learning systems. The third group, consisting of Bosco, Poletto, and Sanguinetti, uses a combination of computational linguistic techniques and data visualization tools to detect for hate speech on Twitter data. These authors work closely together due to their nationality and the proximity of their institutions. The fourth cluster, consisting of Kumar, Roy, and Benamara, explores the use of machine and DL techniques, including BERT, to detect hate speech and offensive content in Dravidian languages.

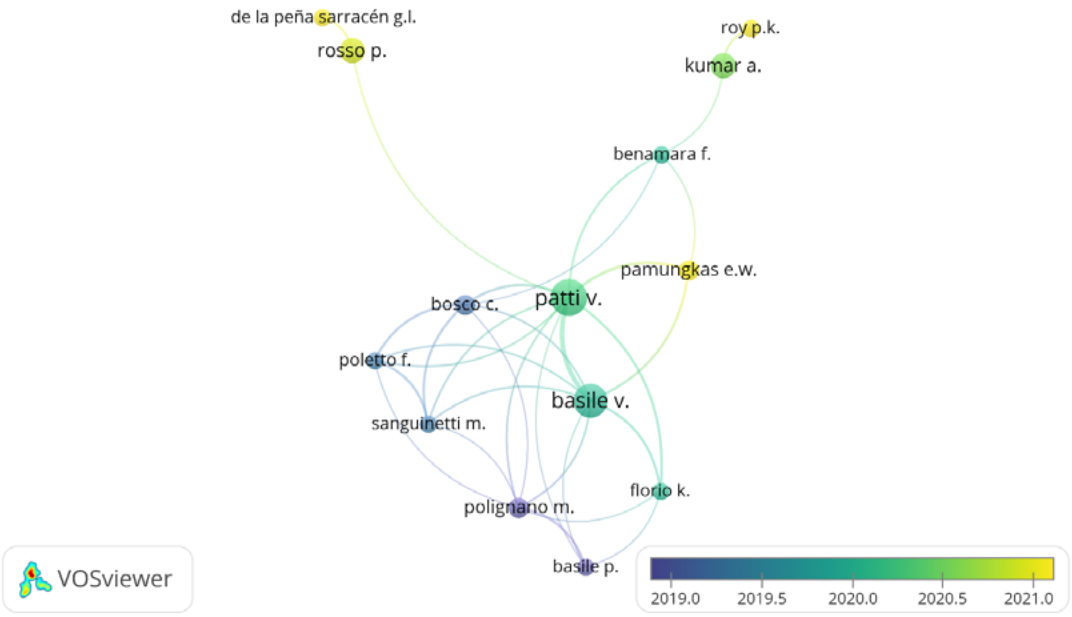

Figure 8. Collaboration among authors via network visualization

Source: Own elaboration.

The last one is constituted by Rosso and De la Peña Sarracén, and focuses on virtual hate speech detection in contexts with limited labeled data, using a convolutional neural network (HaGNN) for text classification. The authors of the last two clusters have the most recent works and highlight the current trends and challenges in the field.

If we focus on the authors' countries of origin, and set a minimum threshold of 10 publications per country, without establishing a minimum number of citations for a country, only 19 of the 79 countries that meet this criterion with each other. The 19 countries are divided into five clusters, represented by different colors (figure 9). Indian researchers collaborate with countries that belong to other clusters. From the collaboration map it can be seen that India, Sri Lanka, Taiwan, and Bangladesh are making significant contributions to the of hate speech in Dravidian languages. This highlights a recent trend in the use of machine and DL techniques and BERT for this purpose (Roy et al., 2022).

While much of the research in this area has focused on the detection of hate speech in English, the map shows collaboration between authors from Indian and German institutions. The HASOC method, which aims to provide a platform for developing and optimizing hate speech detection algorithms for Hindi, German, and English, is also highlighted (Mandl et al., 2020).

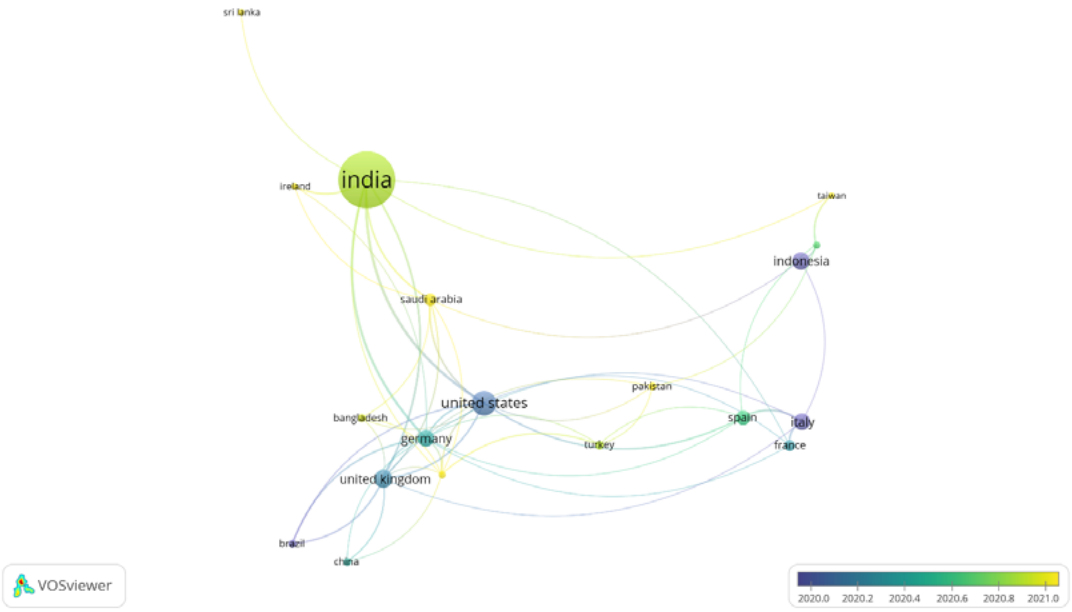

Figure 9 Collaboration among countries via network visualization (No. of publications)

Source: Own elaboration.

Considering institutions as the unit of analysis in a co-authorship analysis reveals organizations that exhibit some degree of collaboration among their authors. In this case, we set as a minimum criterion that each organization must have published at least two articles and received at least one citation.

This criterion is met by 47 organizations, but only three of them show collaboration between their authors: Dhirubhai Ambani I. I. (D.A.I.I.) and Communication Technology (Gandhinagar, India), LDRP Institute of Technology and Research (Gandhinagar, India), and University of Hildesheim (Germany). Authors from D.A.I.I and Communication Technology (Gandhinagar, India), LDRP Institute of Technology and Research (Gandhinagar, India), and University of Hildesheim (Germany) collaborate to analyze data posted on Facebook and Twitter. They use various classifiers, such as SVM and logistic regression, as well as DL models based on CNN and BERT to classify aggression. They also use ICHCL and HASOC methods to identify and filter hate speech by classifying messages in mixed-code languages. These institutions present strong collaboration based on their studies in this field (Mandl et al., 2020; Modha et al., 2020, 2022; Satapara et al., 2021).

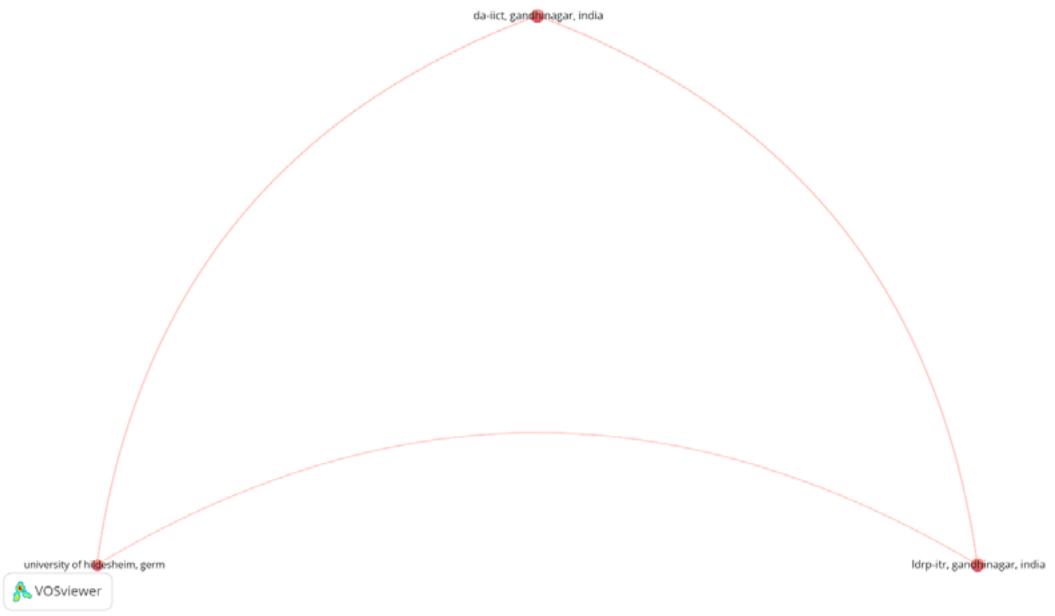

Figure 10. Collaboration among organizations via network visualization (No. of publications)

Source: Own elaboration.

Analysis of joint citation references (co-citation)

Figure 11 shows the map of the co-citation network of the analyzed bibliography. In our study, we collected 16,973 citations, and the minimum threshold for analysis was set at 17 citations and eight articles that are related at least once. The larger the dot, the higher the citation frequency, and the thicker the line between two dots, the closer the relationship, as reflected in the number of co-cited links in the reference (Boyack & Klavans, 2010). The references are grouped in two different clusters, with different colors to represent different topics within the research area.

The red cluster (Waseem & Hovy, 2016; Nobata et al., 2016; Davidson et al., 2017; Schmidt & Wiegand, 2017; Badjatiya et al., 2017) highlights the need to identify and classify the type of hate speech (sexist, xenophobic, etc.) for which they propose different methods that can be combined such as deep neural networks, convolutional networks, and long-term memory networks, DL or semantic word embeddings.

From the green cluster (Bojanowski et al., 2017; Watanabe et al., 2018; Devlin et al., 2019), an area of analysis emerges consisting of the use of algorithms of different types applicable to different languages for hate speech detection.

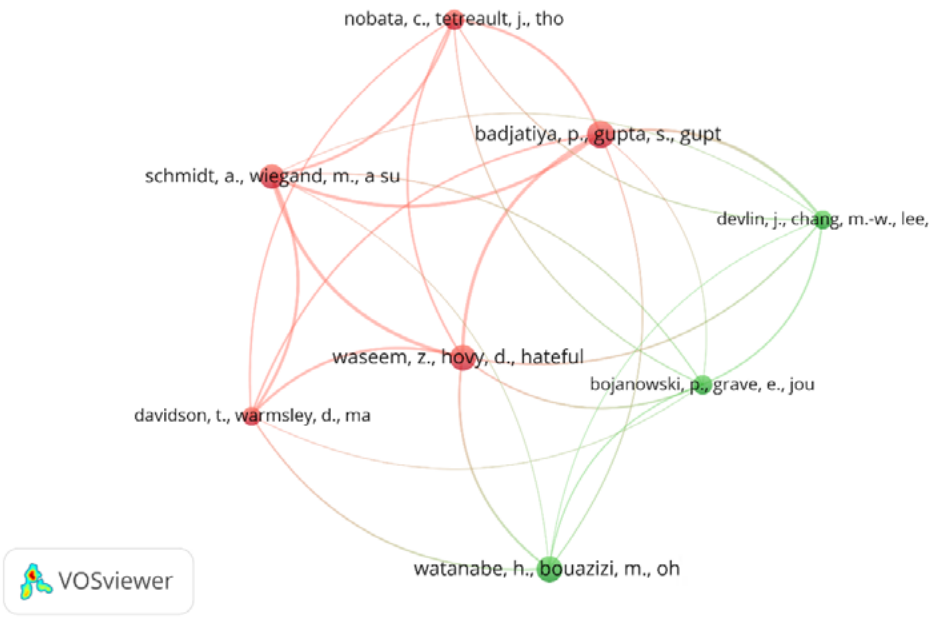

Figure 11. Co-citations of references

Source: Own elaboration.

Keyword co-occurrence analysis

A total of 1,850 index-keywords were extracted. To make the visualization more understandable, a minimum frequency of 10 was set for the keywords. A keyword-thesaurus was created to group keywords that refer to the same concept. Analysis using VOSviewer resulted in four major groups of keywords. The links between the keywords indicate their co-occurrence relationship, and the color of the nodes represents to which each keyword is assigned. The size of the labels and the diameter of the circles indicate the frequency and strength of the links between the keywords. The four identified topics reveal important areas of online hate research.

Red and yellow clusters correspond more to theoretical research topics related to the analysis of discourse and language through identification tasks, semantic analysis, and classification, as suggested by the most relevant words included in them: speech classification, semantics, language detection/model, or decision trees.

Blue and green clusters represent research topics that are more practical in nature. They refer to the automatic detection of hate speech in social media, by applying certain techniques as suggested by the main words included in them: automatic detection, DL, deep neural networks, LSTM, Twitter or Facebook.

If we focus on the keyword "automatic detection" and its relationship to "social media," we can see that the surface of an imaginary circle drawn between the two words would highlight the trendy and emerging techniques for detecting hate speech on social networks.

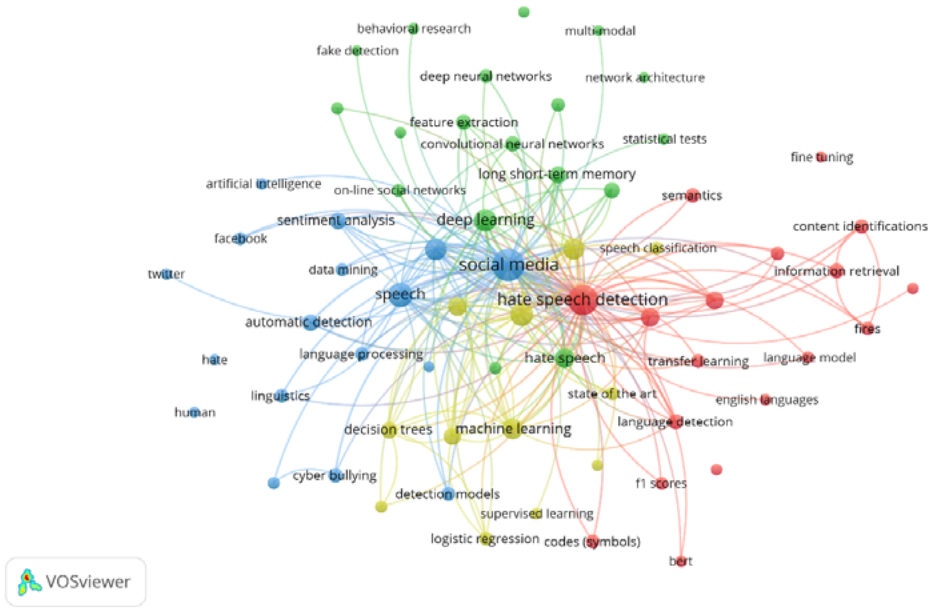

Figure 12. Network co-occurrence index-keywords

Source: Own elaboration.

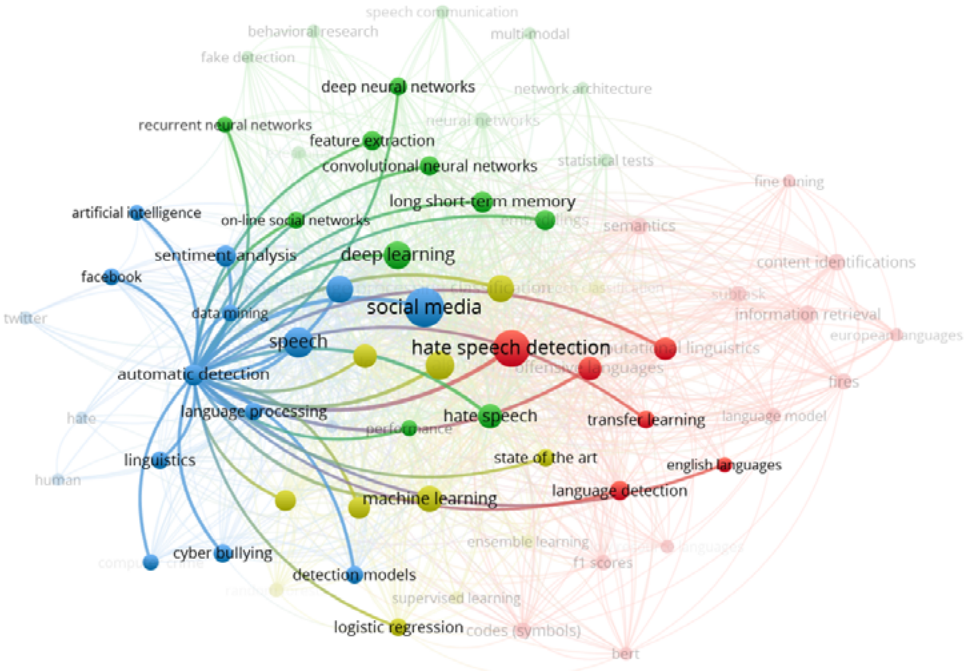

Figure 13. Network co-occurrence keywords focused in automatic detection

Source: Own elaboration.

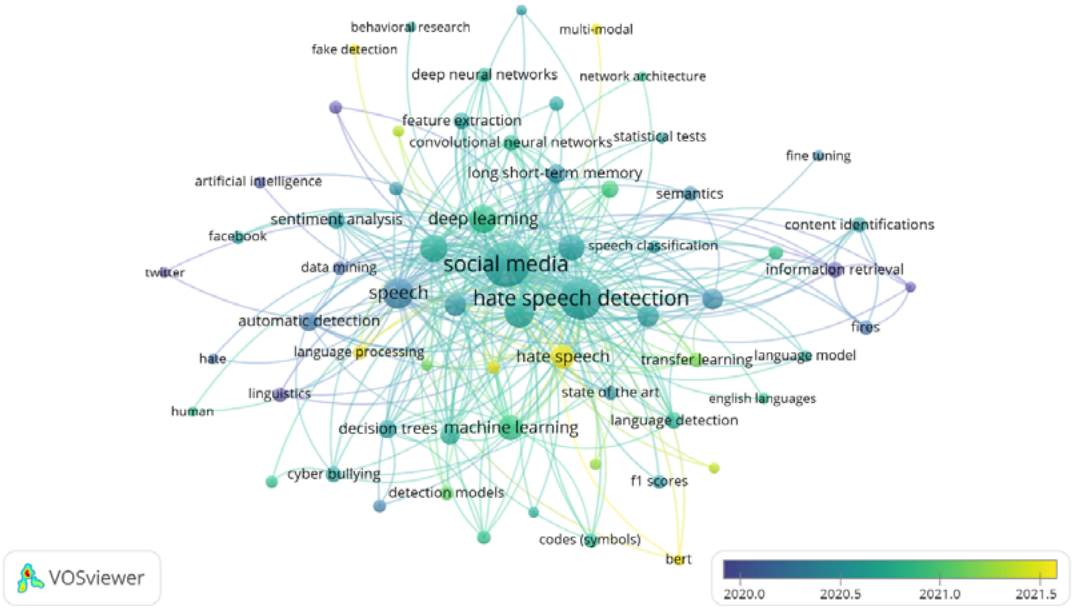

Figure 14. Overlay co-occurrence index keywords

Source: Own elaboration.

These techniques include ML (Sindhu et al. 2020; Saeed et al., 2021; Mutanga et al., 2022), various DL models (Batani et al., 2022) or combinations thereof, convolutional neural networks (CNN) (Elouali et al., 2020; Alotaibi et al., 2021), LSTM (Salim & Suhartono, 2021; Xiang et al., 2021), support vector machines (Liyanage & Jayakumar, 2021; Boulouard et al., 2022), learning algorithms (Putri et al., 2020; Baydogan & Alatas, 2021), and NLPS (Gongane et al., 2022; Li et al., 2022), to improve performance, e.g., through data augmentation and meta-learning in scenarios with scarce or low quality data (Hedderich et al., 2021).

The analysis of the publications' temporal evolution shows that the recent trend in the research in the analyzed field is focused on the detection of fake information in social networks (Bailurkar & Raul,2021; Gongane et al.,2022), and on the application of the BERT model (Bhawal et al.,2021; Roy et al., 2022).

Figure 15 shows a graph created from the index keywords-SCOPUS of the publications used in our analysis, with a minimum occurrence of 10. In general, each point has a color indicating the density of the element. When visualizing element density, the colors can be blue, yellow, orange, and red (from lowest to highest abundance or density). The map shows a structure that resembles concentric rings. The core (red color) is formed by the research topic (hate speech detection and social media). The following rings show the most commonly used techniques for hate speech detection: orange (DL, NLPS, learning algorithms), yellow (text processing, classification, ML), green (LSTM, embeddings, CNN, or sentiment analysis), blue (other techniques).

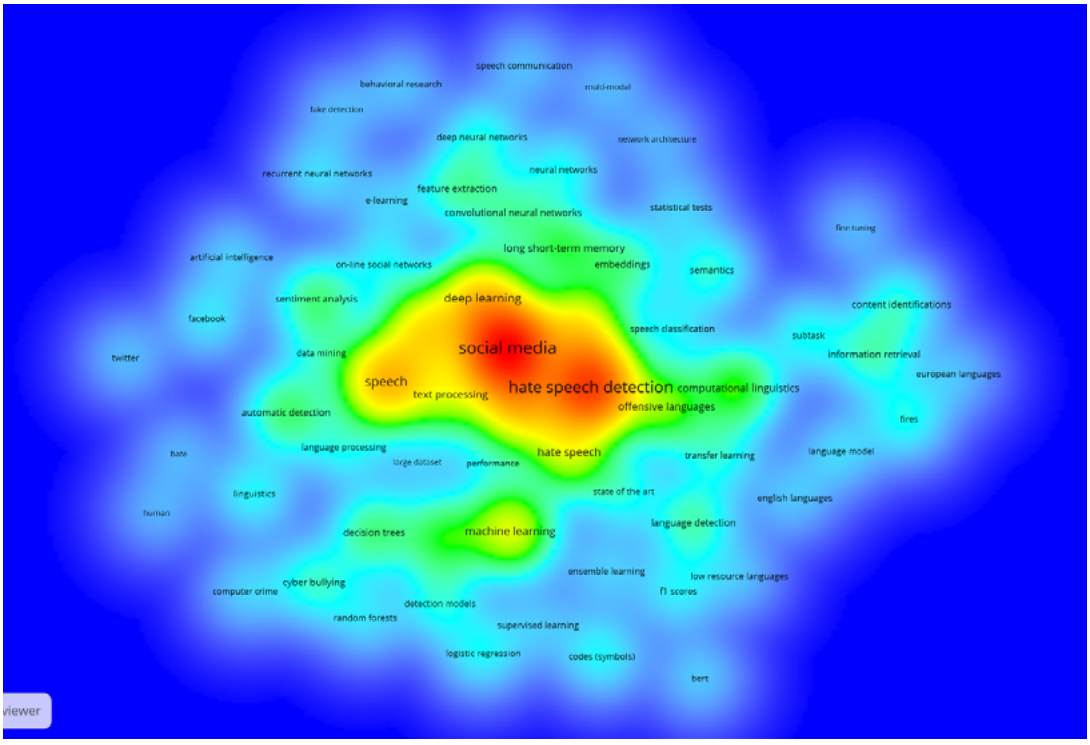

Figure 15. Density visualization co-occurrence keywords

Source: Own elaboration.

DISCUSSION

The increase in research on hate speech and social media shows that the topic is of great interest to the scientific community. This trend is fueled by the rapid development of information and communication technologies and the increasing number of social media users worldwide, as more than half of the world's population using these platforms (Galeano, 2021). The abundance of hate speech on social media has led to the need to control it (Gascón, 2019; European Commission, 2020). The data obtained from Scopus show a positive trend in the number of relevant publications since 2016, although a decrease was observed from 2021 to 2022 (Ramírez-García et al., 2022). This issue is relatively new and has been little researched. It is the result of increasing concern raised by international organizations such as the Movement Against Intolerance (2015) and European Commission Against Racism and Intolerance (2015) about the growing problem of hate speech.

The Computer Science field is the most active when it comes to detect hate speech in social media (Ramírez-García et al., 2022). Other disciplines such as linguistics, politics, education, and others also contribute, reflecting the interdisciplinary nature of the research area (Gangurde et al., 2022).

Researchers from the United States, India, the United Kingdom, and Germany frequently collaborate in this area of research (Gangurde et al., 2022).

As far as social media are concerned, Twitter is considered the most widespread and relevant platform for spreading hate speech. There is a close collaboration between researchers from the Dhirubhai Ambani I.I. and Communication Technology (Gandhinagar, India), LDRP Institute of Technology and Research (Gandhinagar, India), and University of Hildesheim (Germany). The result is a line of research focused on analyzing data posted on Facebook and Twitter using plugins that process both English and Hindi data in combination with code. To classify hate speech, the researchers use various classifiers including Support Vector Machine (SVM), logistic regression, Convolution Neural Network (CNN)-based DL models, attention-based models, and the pre-trained linguistic model BERT.

Research on the detection of hate speech in social media in India is prominent, with a high number of published articles, prolific organizations, and influential authors.

Analysis of the temporal trends in the use of various control mechanisms has been applied in a novel way in this field of research, and a positive trend has been observed.

Nevertheless, there are still some challenges that require attention and could provide opportunities for future research. These include:

- Implement automatic detection of hate speech in closed systems, as individuals may try to evade detection if they are aware of being monitored.
- The problem of detecting hate speech or insults in Hindi, which requires handling code mixing between Hindi and English.
- Development of new approaches, methods, or algorithms for generating data in low or poor information quality scenarios.
- There is a lack of studies that address the analysis and handling of multimedia content that spreads hate (images, videos, audios, etc.).
- The comparison between different approaches provides an interesting approach, to help professionals choose the right tool for their task.
- Evaluate the long-term trend in the number of publications to determine whether the decrease from 2021 to 2022 is a temporary phenomenon or the result of effective control mechanisms.

CONCLUSION

The use of social networks is becoming increasingly important in our daily lives and interactions with others, making the detection of hate speech detection on these platforms crucial. Their automatic detection by ML is a promising approach to stop their spread Classification of abusive language with a model based exclusively on textual data has been shown to have limited performance due to the complexity and diversity of speech.

The study's results show that using bibliometric tools such as VOSviewer and Refworks can provide a comprehensive understanding of the field through by creating maps displaying co-occurrence, co-citations, and density distribution, thus providing useful insights and facilitating the work of future researchers in the field of automatic detection of hate speech in social networks by highlighting emerging trends, prominent authors and publications, important sources, keywords, and research impact. Our findings can guide future researchers and highlight areas for further study.

Our study highlights the multifaceted and multidisciplinary approach to detecting hate speech in social media, that combines linguistic, social behavior, and psychological analyses. The most commonly used techniques are ML, DL, and NLP. Of these, ML is the most widely used method that incorporates that incorporates NLP and uses pre-trained models such as BERT to achieve effective results in detecting hate speech.

This study has limitations as it only considers publications from the Scopus database from 2016 to 2022, and does not include other m databases (WoS, Springer, or ScienceDirect).

The visualizations presented in this paper were created using the VOSviewer software and saved as images. This means that some details may not be visible in the figures as they were taken as screenshots.

REFERENCES

Al-Hassan, A. & Al-Dossari, H. (2022). Detection of hate speech in Arabic tweets using deep learning. *Multimedia Systems*, *28*, 1963–1974. https://doi.org/10.1007/s00530-020-00742-w

Albadi, N., Kurdi, M., & Mishra, S. (2018). Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM) (pp. 69-76). https://doi.org/10.1109/ASONAM.2018.8508247

Alotaibi, M., Alotaibi, B., & Razaque, A. (2021). A Multichannel Deep Learning Framework for Cyberbullying Detection on Social Media. *Electronics*, *10*(21), 2664. https://doi.org/10.3390/electronics10212664

Alrashidi, B., Jamal, A., Khan, I., & Alkhathlan, A. (2022). A review on abusive content automatic detection: approaches, challenges and opportunities. *PeerJ Computer Science*, *8*, e1142. https://doi.org/10.7717/peerj-cs.1142

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. In R. Barret & R. Cummings (Chairs), *WWW '17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 759-760). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3041021.3054223

Bailurkar, R. & Raul, N. (2021). Detecting Bots to Distinguish Hate Speech on Social Media. In *2021 12th International Conference on Computing Communication and Networking Technologies* (pp. 1-5). IEEE. https://doi.org/10.1109/ICCCNT51525.2021.9579883

Batani, J., Mbunge, E., Muchemwa, B., Gaobotse, G., Gurajena, C., Fashoto, S., Kavu, T., & Dandajena, K. (2022). A Review of Deep Learning Models for Detecting Cyberbullying on Social Media Networks. In R. Silhavy (Ed.), *Cybernetics Perspectives in Systems - Proceedings of 11th Computer Science On-line Conference, CSOC 2022* (vol. 3) (pp. 528-550). Springer Science. https://doi.org/10.1007/978-3-031-09073-8_46

Baydogan, C. & Alatas, B. (2021). Metaheuristic Ant Lion and Moth Flame Optimization-Based Novel Approach for Automatic Detection of Hate Speech in Online Social Networks. In *IEEE Access*, *9*, 110047-110062. https://doi.org/10.1109/ACCESS.2021.3102277

Ben-David, A. & Matamoros Fernández, A. (2016). Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain. *International Journal of Communication*, *10*, 1167–1193. https://ijoc.org/index.php/ijoc/article/view/3697

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection. In M. Nissim, V. Patti, B. Plank, C. Wagner (Eds.), *Proceedings of the 2nd Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media* (pp. 36-41). Association for Computational Linguistics. https://aclanthology.org/W18-1105

Bojanowski, P., Grave, E., Joulin, A., & Tomas Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*,135–146. https://doi.org/10.1162/tacl_a_00051

Boulouard, Z., Ouaissa, M., & Ouaissa, M. (2022). Machine Learning for Hate Speech Detection in Arabic Social Media. In M. Ouaissa, Z. Boulouard, M. Ouaissa, B. Guermah, (Eds.), *Computational Intelligence in Recent Communication Networks* (pp. 147-162). Springer. https://doi.org/10.1007/978-3-030-77185-0_10

Boyack, K. W. & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389-2404. https://doi.org/10.1002/asi.21419

Bhawal, S., Roy, P. K., & Kumar, A. (2021). Hate Speech and Offensive Language Identification on Multilingual Code-Mixed Text Using BERT. *CEUR Workshop Proceedings, 3159*, 615-624. https://ceur-ws.org/Vol-3159/

Burnap, P. & Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*, 11. https://doi.org/10.1140/epjds/s13688-016-0072-6

Córdoba-Cely, C., Alpiste, F., Londoño, F., & Monguet, J. (2012). Análisis de cocitación de autor en el modelo de aceptación tecnológico, 2005-2010 (Author Co-citation Analysis of the Technology Acceptance Model, 2005-2010). *Revista Española De Documentación Científica*, *35*(2), 238–261. https://doi.org/10.3989/redc.2012.2.864

Dadvar, M., Trieschnigg, D., Ordelman, R., & De Jong. F. (2015). Improving Cyberbullying Detection With User Context. In P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, & E. Yilmaz (Eds.), *Advances in Information Retrieval. ECIR 2013. Lecture Notes in Computer Science* (vol. 7814) (pp. 693-696). Springer.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512-515. https://doi.org/10.1609/icwsm.v11i1.14955

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. *Proceedings of the First Italian Conference on Cybersecurity, 1816*, 86-95.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805

Dewani, A., Memon, M. A., & Bhatti, S. (2021). Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data. *Journal of Big Data*, *8*, 160. https://doi.org/10.1186/s40537-021-00550-7

Elouali, A., Elberrichi, Z., & Elouali, N. (2020). Hate Speech Detection on Multilingual Twitter Using Convolutional Neural Networks. *Revue d'Intelligence Artificielle*, *34*, 81-88. https://doi.org/10.18280/ria.340111

European Commission. (2020, June 22). *El código de conducta de la UE para la lucha contra la incitación ilegal al odio en Internet* (Commission publishes EU Code of Conduct on countering illegal hate speech online continues to deliver results) (press release IP/20/1134). https://ec.europa.eu/commission/presscorner/detail/es/IP_20_1134

ECRI General Policy Recommendation No. 15 on Combating Hate Speech and Explanatory Memorandum of December 8, 2015. Strasbourg, France, March 21, 2016.

Fernández, M., Valbuena, C., & Caro, C. (2015). *Evolución del racismo, la xenofobia y otras formas conexas de intolerancia en España* (Evolution of racism, xenophobia, and other intolerance-related forms in Spain). Subdirección General de Información Administrativa y Publicaciones. https://inclusion.seg-social.es/oberaxe/es/publicaciones/documentos/documento_0089.htm

Fersini, E., Nozza, D., & Boifava, G. (2020). Profiling Italian Misogynist: An Empirical Study. In J. Monti, V. Basile, M. P. Di Buono, R. Manna, A. Pascucci, & S. Tonelli (Eds.), *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language* (pp. 9-13). ELRA. https://aclanthology.org/volumes/2020.restup-1/

Fortuna, P. & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR), 51*(4), 1-30. https://doi.org/10.1145/3232676

Galeano, S. (2021). Cuáles son las redes sociales con más usuarios del mundo (2023) (Which are the social networks with the most users in the world? (2023)). *Marketing Ecommerce*. https://marketing4ecommerce.net/cuales-redes-sociales-con-mas-usuarios-mundo-ranking/

Gascón, A. (2019). La lucha contra el discurso del odio en línea en la Unión Europea y los intermediarios de Internet (Fighting online hate speech in the European Union and Internet intermediaries). In Z. Combalía, M. P. Diago, & A. González-Varas (Coords.), *Libertad de expresión y discurso de odio por motivos religiosos* (Freedom of speech and religiously motivated hate speech) (pp. 64-86). Ediciones del Licregdi.

Giumetti, G.W., Robin, M., & Kowalski, R.M. (2022). Cyberbullying via social media and well-being. *Current Opinion in Psychology*, *45*, 101314. https://doi.org/10.1016/j.copsyc.2022.101314

Glänzel, W. & Schubert, A. (2004). Analysing Scientific Networks Through Co-Authorship. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research*. Springer. https://doi.org/10.1007/1-4020-2755-9_12

Gangurde, A., Mankar, P., Chaudhari, D., & Pawar, A. (2022). A Systematic Bibliometric Analysis of Hate Speech Detection on Social Media Sites? *Journal of Scientometric Research*, *11*(1), 100-111.

Gongane, V. U., Munot, M. V., & Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: Current status and future directions. *Social Network Analysis and Mining*, *12*, 129. https://doi.org/10.1007/s13278-022-00951-3

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., Klakow, D. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2545–2568). https://doi.org/10.48550/arXiv.2010.12309

Kanan, T., Aldaaja, A., Hawashin, B. (2020). Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in Arabic social media contents. *The Journal of Internet Technology*, *21*(5),1409-1421. https://jit.ndhu.edu.tw/article/view/2376

Li, C. T., Ku, L. W., Tsai, Y. C., & Wang, W. Y. (2022). SocialNLP'22: 10th international workshop on natural language processing for social media. In F. Laforest, R. Troncy, L. Médini, & I. Herman (Eds.), *Companion Proceedings of the Web Conference 2022* (pp. 849-851). ACM. https://doi.org/10.1145/3487553.3524876

Liyanage, O. & Jayakumar, K. (2021). Hate Speech Detection in Sinhala-English Code-Mixed Language. In *Proceedings of the 21st International Conference on Advances in ICT for Emerging Regions, ICter* (pp. 225-230). IEEE. https://doi.org/10.1109/ICter53630.2021.9774816

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS ONE, 14*(8). https://doi.org/10.1371/journal.pone.0221152

Mandl, T., Modha S., Kumar M, A., & Chakravarthi, B.J. (2020). Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *FIRE '20: Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation* (pp. 29-32). ACM. https://doi.org/10.1145/3441501.3441517

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado-López, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, *126*, 871-906. https://doi.org/10.1007/s11192-020-03690-4

Mishra, R. (2021). Are We Doing Enough? A Bibliometric Analysis of Hate Speech Research in the Selected Database of Scopus. *Library Philosophy and Practice*, 5140. https://digitalcommons.unl.edu/libphilprac/5140/

Modha, S., Majumder, P., & Mandl, T. (2022). An empirical evaluation of text representation schemes to filter the social media stream. *Journal of Experimental and Theoretical Artificial Intelligence*, *34*(3), 499-525. https://doi.org/10.1080/0952813X.2021.1907792

Modha, S., Majumder, P., Mandl, T., & Mandalia, C. (2020). Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance. *Expert Systems with Applications*, *161*, 113725. https://doi.org/10.1016/j.eswa.2020.113725

Mondal, M., Araújo Silva, L., & Benevenuto, F. (2017). *A measurement study of hate speech in social media*. In P. Dolog & P. Vojtas (Chairs), *HT '17: Proceedings of the 28th ACM Conference on Hypertext and Social Media* (pp. 85-94). ACM. https://doi.org/10.1145/3078714.3078723

Movement Against Intolerance. (2015). *Informe Raxen. Racismo, Xenofobia, Antisemitismo, Islamofobia, Neofascismo, Homofobia y otras manifestaciones relacionadas de Intolerancia a través de los hechos. Especial Acción Jurídica contra el Racismo y los Crímenes de Odio.* https://inclusion.seg-social.es/oberaxe/es/publicaciones/documentos/documento_0013.htm

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). *A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media*. In H. Cherifi, S. Gaito, J. Mendes, E. Moro, & L. Rocha (Eds.), *Complex Networks and Their Applications VIII. COMPLEX NETWORKS 2019. Studies in Computational Intelligence* (vol. 881) (pp. 928-940). Springer. https://doi.org/10.1007/978-3-030-36687-2_77

Mutanga, R. T, Naicker, N, & Olugbara, O. O. (2022). Detecting Hate Speech on Twitter Network using Ensemble Machine Learning. *International Journal of Advanced Computer Science and Applications*, *13*(3). https://doi.org/10.14569/IJACSA.2022.0130341

Nandiyanto, A. B. D., Biddinika, M. K., & Triawan, F. (2020). How bibliographic dataset portrays decreasing number of scientific publication from Indonesia. *Indonesian Journal of Science and Technology*, *5*(1), 154-175. https://doi.org/10.17509/ijost.v5i1.22265

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. In J. Bourdeau, J. A. Hendler, & R. Nkambou (Chairs), *WWW '16: Proceedings of the 25th International Conference on World Wide Web* (pp. 145-153). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2872427.2883062

Omar, A. & Hashem, M. E. (2022). An Evaluation of the Automatic Detection of Hate Speech in Social Media Networks. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *13*(2). https://doi.org/10.14569/IJACSA.2022.0130228

Page, M. J., MacKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffman, T. C., Mulrow, C. D., Shamseer, L., Telzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A. Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinnes..., & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ, 372*(71). https://doi.org/10.1136/bmj.n71

Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, *48*, 4730-4742. https://doi.org/10.1007/s10489-018-1242-y

Putri, T. T. A., Sriadhi, S., Sari, R. D., Rahmadani, R., & Hutahaean, H. D. (2020). A comparison of classification algorithms for hate speech detection. *IOP Conference Series: Materials Science and Engineering*, *830*(3). https://doi.org/10.1088/1757-899X/830/3/032006

Ramírez-García, A., González-Molina, A., Gutiérrez-Arenas, M., & Moyano-Pacheco, M. (2022). Interdisciplinariedad de la producción científica sobre el discurso del odio y las redes sociales: Un análisis bibliométrico (Interdisciplinarity of scientific production on hate speech and social media: A bibliometric analysis). *Comunicar*, *72*, 129-140. https://doi.org/10.3916/C72-2022-10

Risch, J. & Krestel, R. (2018). Aggression Identification Using Deep Learning and Data Augmentation. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)* (pp. 150-158). Association for Computational Linguistics. https://aclanthology.org/W18-4418/

Romero, L. & Portillo-Salido, E. (2019). Trends in Sigma-1 Receptor Research: A 25-year Bibliometric Analysis. *Frontiers in Pharmacology, 10*. https://doi.org/10.3389/fphar.2019.00564

Roy, P. K., Bhawal, S., & Subalalitha, Ch. N. (2022). Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. *Computer Speech & Language, 75*,101386. https://doi.org/10.1016/j.csl.2022.101386

Saeed, F., Al-Sarem, M., & Alromema, W. (2021). Tuning Hyper-Parameters of Machine Learning Methods for Improving the Detection of Hate Speech. In F. Saeed, T. Al-Hadhrami, F. Mohammed, & E. Mohammed (Eds.), *Advances on Smart and Soft Computing. Advances in Intelligent Systems and Computing* (vol. 1188) (pp. 71-78). Springer. https://doi.org/10.1007/978-981-15-6048-4_7

Salim, C. E. R. & Suhartono, D. (2021). Long Short-Term Memory for Hate Speech and Abusive Language Detection on Indonesian Youtube Comment Section. In H. Lin (Ed.), *Proceedings of the 2021 11th International Workshop on Computer Science and Engineering* (pp. 193-200). https://doi.org/10.18178/wcse.2021.06.029

Satapara, S., Modha, S., Mandl, T., Madhu, H., & Majumder, P. (2021). Overview Of the HASOC Subtrack At FIRE 2021: Conversational Hate Speech Detection in Code-Mixed Language. In P. Mehta, T. Mandl, P. Majumder, & M. Mitra (Eds.), *FIRE-WN 2021: FIRE 2021 working notes* (pp. 20-31). RWTH.

Schmidt, A. & Wiegand, A. (2017). A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1–10). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1101

Sellars, A. F., (2016). Defining Hate Speech. *Public Law Research*, 16-48. https://doi.org/10.2139/ssrn.2882244

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). *Analyzing the Targets of Hate in Online Social Media. Proceedings of the International AAAI Conference on Web and Social Media*, *10*(1), 687-690. https://doi.org/10.1609/icwsm.v10i1.14811

Sindhu, A., Sarang, S., Zahid, H. K, Zafar, A., Sajid, K. & Ghulam, M. (2020). Automatic Hate Speech Detection using Machine Learning: A Comparative Study. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *11*(8). https://doi.org/10.14569/IJACSA.2020.0110861

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayer, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, *126*, 5113-5142. https://doi.org/10.1007/s11192-021-03948-5

Strossen, N. (2016). Freedom of Speech and Equality: Do We Have to Choose? *Journal of Law and Policy*, *25*(1), 185–225.

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, *126*, 157-179. https://doi.org/10.1007/s11192-020-03737-6

UNESCO. (2021). *Addressing Hate Speech on Social Media: Contemporary Challenges*. https://unesdoc.unesco.org/ark:/48223/pf0000379177

United Nations. (2019). *UN Strategy and Plan of Action on Hate Speech*. https://www.un.org/en/hate-speech

Van Eck, N. J. & Waltman, L. (2020). *VOSviewer Manual*. Universiteit Leiden. https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.8.pdf

Vargas-Quesada, B., Chinchilla-Rodríguez, Z., & Rodríguez, N. (2017). Identification and Visualization of the Intellectual Structure in Graphene Research. *Frontiers in Research Metrics and Analytics*, *2*. https://doi.org/10.3389/frma.2017.00007

Waseem, Z. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science* (pp. 138-142). Association for Computational Linguistics.

Waseem, Z. & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop* (pp. 88–93). Association for Computational Linguistics.

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate Speech on Twitter A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. *IEEE Access, 6*, 13825-13835. https://doi.org/10.1109/ACCESS.2018.2806394

Xiang, K., Zhang, Z., Yu, Y., San Lucas, L., Amin, M. R., & Li, Y. (2021). Identification of Hate Tweets: Which Words Matter the Most? In C. Stephanidis, M. Kurosu, J. Y. C. Chen, G. Fragomeni, N. Streitz, S. Konomi, H. Degen, & S. Ntoa (Eds.), *HCI International 2021 - Late Breaking Papers: Multimodality, eXtended Reality, and Artificial Intelligence* (pp. 586-598). Springer. https://doi.org/10.1007/978-3-030-90963-5_44

Zamora-Bonilla, J. & González de Prado Salas, J. (2014). Un análisis inferencialista de la co-autoría de artículos científicos (an inferentialist conception regarding the co-authorship of scientific papers). *Revista Española de Documentación Científica*, *37*(4), e064. https://doi.org/10.3989/redc.2014.4.1145

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (pp. 1415-1420). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1902.09666

ABOUT THE AUTHORS

ANA M. SÁNCHEZ-SÁNCHEZ, associate professor in the Department of Economics, Quantitative Methods, and Economic History at the Universidad Pablo de Olavide. Ph.D in Business Administration and Management from the Universidad Pablo de Olavide. Her research interests include poverty indicators, tourism economics, sustainable development, tourist behavior, and labor tourism. She has published on impact evaluation of economy and tourism journals. She is a member of the Multidisciplinary Statistical and Demoscopic Studies research group.

http://orcid.org/0000-0002-6591-954X

DAVID RUIZ-MUÑOZ, internal auditor at the *Junta of Andalusia*. He has been a lecturer in the Department of Economics, Quantitative Methods, and Economic History and the Department of Financial Economics and Accounting at the Universidad Pablo de Olavide. Ph.D. in Business Administration and Management from the Universidad Pablo de Olavide. He has published on impact evaluation of economics and sociology journals.

https://orcid.org/0000-0003-4538-7774

FRANCISCA J. SÁNCHEZ-SÁNCHEZ, associate professor in the Department of Economics, Quantitative Methods, and Economic History at the Universidad Pablo de Olavide. Ph.D. in Business Administration and Management from the Universidad Pablo de Olavide. Her research interests include the study of models, specifically applying the Multivariate Analysis and DEA methodology. This group of contributions includes applied papers, which enable her to participate in works with diverse topics. She has published on impact evaluation of economy and tourism journals.

http://orcid.org/0000-0001-5325-3667